A network management system for characterizing high-rate, large-sized flows

Tian Jin*, Chris Tracy†, Malathi Veeraraghavan*, Zhengyang Liu *
* University of Virginia
Charlottesville, VA 22904–4743
Email: {tj3sr,mvee,zl4ef}@virginia.edu
† Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory
Berkeley, CA 94720
Email: ctracy@es.net

***Abstract*—High-rate, large-sized (α) flows are of interest to providers for various reasons. For example, they have the potential to degrade service quality for real-time flows. Also, users who generate these flows are sensitive to performance measures such as throughput and throughput variance. In this paper, we describe an algorithm for characterizing the size, duration, average rate, and frequency of such flows, from NetFlow reports that are already collected from routers by most providers. The algorithm was validated using independently collected usage logs from application servers. This algorithm can be used in a network management system for providers interested in these types of flows, such as research-and-education network providers whose customers move large scientific datasets. We executed the algorithm on actual NetFlow reports from 4 ESnet routers collected over a 7-month period. Flows moving datasets as large as 811 GB and at rates as high as 5.7 Gbps were observed. Some source-destination pairs were found to repeatedly create α flows. An analysis of the rates of the 1596 repeated α flows created by one pair showed considerable variance, with minimum rate of 100 Mbps, maximum rate of 536 Mbps, and a coefficient of variation of 30%.**

I. Introduction

Problem statement: Develop an algorithm for characterizing high-rate large-sized flows (which we refer to as α flows [1]) for use in a network management system.

Motivation: Core research-and-education networks (RENs), such as US Department of Energy (DOE)'s Energy Sciences Network (ESNet) [2] and Internet2 [3], offer connectivity services to DOE laboratories and universities, respectively. For example, ESnet4[1] had about 77 routers, with core routers in major US cities, provider-edge routers in several of its 43 customer sites, and three metropolitan-area ring networks. Scientific research conducted at the DOE laboratories and universities often entails executing highly parallelized applications on supercomputing platforms, and then transferring large datasets to local storage clusters [4]. We analyzed GridFTP (a high-speed file-transfer application used by the scientific community [5]) usage logs from operational data transfer nodes deployed at three supercomputing centers and found transfers with rates as high as 4 Gbps, and sizes in the hundreds of GB range [6].

Network operators are interested in characterizing such high-rate large-sized flows traversing their network for various applications. A few examples are as follows. *First*, since such flows cause traffic burstiness, their presence can cause increased delay and jitter in real-time flows. Therefore, there is interest in traffic engineering systems that can identify α flows at their ingress routers, and isolate them from general-purpose flows using a separate service class or virtual circuits [7]. *Second*, while REN peerings (e.g., with GEANT, SINET, CERNET, KREONET, ERNET) are usually the preferred routes for inter-domain traffic within the scientific community, sometimes these α flows moving large scientific datasets appear on the commercial peering links. Such events occur due to BGP misconfigurations. A network management system that characterizes α flows can assist providers in finding such misconfigurations. A *third* use of a system that characterizes α flows will enable providers to assist their customers in determining causes of poor performance. For example, if a user experiences high throughput variance as determined by our network management system, PerfSONAR [8] can be used to help pinpoint the source of the problem. In summary, there is interest from operators and in the research community as shown in Section II to characterize α flows.

Solution approach: NetFlow, a built-in capability of (provider) IP routers, samples packets and periodically exports reports to an external server. We developed an algorithm for combining information from multiple NetFlow reports to determine the size, duration, and average rate of high-rate, large-sized flows. Given the low NetFlow packet sampling rates (e.g., 1-in-1000)[2], the algorithm needs to be validated. We conducted a validation exercise by procuring GridFTP usage logs from a supercomputing center that is directly connected to ESnet, and NetFlow reports from the corresponding ESnet router. The GridFTP usage logs provide file transfer sizes/durations. These were matched with the flow characteristics determined by executing our algorithm on the ESnet NetFlow reports. The algorithm was then applied to

[1]ESnet4 was recently upgraded to ESnet5 using 100 Gbps links, but since the traffic analyzed was from ESnet4, we describe this network briefly.

[2]On high-speed core-network links, higher sampling rates are impractical.

characterize high-rate large-sized flows observed at four ESnet routers.

Findings/contributions: Our *first* finding is that in spite of low packet sampling rates, the size, duration, and rate of α flows can be accurately estimated from NetFlow reports. *Next*, by executing the algorithm on NetFlow reports gathered from four ESnet routers over a 7-month period, we found flow sizes as large as 811 GB and average rates as high as 5.7 Gbps (backbone link rate in ESnet4 was 10 Gbps). A *comparison* of flow characteristics at different types of routers showed that there were more α flows in the download direction from DOE labs than in the upload direction to DOE labs (which is consistent with the fact that most university scientists use the DOE supercomputing centers to run their applications and then download datasets from these centers). To study *persistency*, we determined the number of flows created by each source-destination IP address pair. The maximum number of α flows (flows that exceed 5GB in size and 100Mbps in rate) for a single source-destination pair was 1596, of which 75% experienced less than 167 Mbps while the highest rate was 536 Mbps. Such information is useful for initiating diagnostics to improve performance.

Novelty and significance: While size/rate characterization for all flow types is challenging because of the low packet sampling rates offered by built-in router features such as NetFlow, our work offers a solution for characterizing size and average rate for a specific subset of flows, i.e., high-rate, large-sized flows. Our validation approach of using operational data from two disparate sources (GridFTP usage logs from file-transfer application servers, and NetFlow reports from ESnet routers) was challenging to execute because of privacy considerations, but it demonstrates the feasibility of validating proposed solutions in an operational context rather than on an experimental testbed. Finally, the significance of the proposed network management system is demonstrated through applications, e.g., providing users and operators information about variance in throughput to help improve performance.

After reviewing related work in Section II, we present our size-rate estimation algorithm in Section III. Algorithm validation is described in Section IV. Section V presents numerical results for high-rate large-sized flows observed at four routers of ESnet, and discusses potential applications for a network management system based on our algorithm. Our conclusions are presented in Section VI.

II. Related work

Kamiyama and Mori propose a short-timeout method to identify high-rate flows [9] and elephant (large) flows [10] with low false-positive and false-negative rates, but not to determine the flow rates or sizes. Zhang, Fang and Zhang [11] proposed a Bayesian single sampling method to identify high-rate flows, but again not to characterize their sizes/rates.

Duffield, Lund and Thorup [12] had a goal of finding information about flows in unsampled packets using information in sampled packets. In contrast, our goal is more specific to characterizing α flows. Given the higher rate of sampling of these flows, our method will result in higher accuracy but is not as general in its scope [12].

There are several papers proposing methods for identifying large flows or high-rate flows with new router hardware. These include ElephantTrap [13], RATE [14], CATE [15], an FPGA-based cache solution [16], and a Grid flow real-time detector for 1 Gbps links [17]. Also Hohn and Veitch [18] proposed a scheme for finding the spectral density, distribution of the number of packets per flow, and showed why alternate sampling techniques were need to obtain this second-order statistic about flows. Given our focus on designing network management systems and not new router hardware, our scheme relies on the built-in NetFlow system supported in most deployed provider routers.

III. Algorithm for estimating flow size/rate

In this section, we provide a brief review of NetFlow, define our terminology and describe the algorithm.

A. NetFlow

NetFlow is a feature that enables IP routers to collect packet samples, and save information on a per-flow basis. The defining attributes of a flow can be configured, e.g., the five tuples {source IP address, destination IP address, source port number, destination port number, protocol type}. For each newly observed flow *F*, NetFlow opens a flow report and stores the arrival time instant of the first observed packet. For every new packet corresponding to flow *F* that is captured by the sampling process, NetFlow adds one to the flow-report packet count and increases the total report size (bytes) by the packet-payload size. It also updates the last-packet timestamp field. At the end of the *active timeout interval* (time since first-packet arrival), *inactive timeout interval* (time since last-packet arrival), or upon observing a TCP FIN or RST segment for flow *F*, the corresponding open NetFlow report is closed. The two timeout intervals are configurable. The closed NetFlow reports are sent by the IP router's `NetFlow exporter` to a `NetFlow collector` (a process running on an external host). In ESnet, the packet sampling rate is 1-in-1000, the active and inactive timeout intervals are 60 sec each, and NetFlow reports are exported every 5 mins.

B. Terminology

Flow: A *flow* is defined to consist of all packets arriving with the same 5-tuple values {source IP address, destination IP address, source port number, destination port number, protocol type} with no consecutive inter-packet gaps greater than some fixed time threshold τ. Inter-packet gaps within the period of a NetFlow report, which are not recorded, are necessarily smaller than the active timeout interval. Therefore the fixed time threshold τ should be at least as large as the NetFlow active timeout interval. The five tuples constitute the *flow IDentifier (flow ID).*

The fixed time threshold phrase is required because a TCP connection can be held open for a long duration, but only carry

packets in intermittent bursts. For example, with HTTP1.1, a TCP connection is held open while a Web client accesses a Web server. If the first downloaded Web page has multiple images located on the same Web server, then each of those images will be downloaded on the same TCP connection. Since the Web client software parses the HTML page and automatically sends out GET requests for the images, these inter-GET time gaps will be short. On the other hand, when human user input (e.g., a mouse click) is required to generate GET requests, there could be large "think-time" gaps.

Multiple sets of GET request bursts (consisting of GET requests generated automatically by the Web client), and their responses, could thus occur on the same TCP connection, and will hence share a flow ID. But packets related to each such set is likely be parsed out as a separate flow by our algorithm given the time threshold phrase in our definition of a *flow*. Effectively, if the time gap between the last-packet timestamp in one NetFlow report r, and the first-packet timestamp in the next NetFlow report with the same flow ID exceeds a threshold, then a flow is said to have terminated with NetFlow report r, and a new flow started with the next NetFlow report.

NetFlow reports: A NetFlow report r is represented as

$$\{\omega_r, f_r, l_r, v_r, o_r\} \tag{1}$$

where ω_r is the (5-tuple) flow identifier, f_r is the Coordinated Universal Time (UTC) timestamp of the first packet in the report, l_r is the UTC timestamp of the last packet in the report, v_r is the number of packets in the report, and o_r is the cumulative number of octets (bytes) in the report.

Types of NetFlow reports: A size threshold H is used to divide a day's set of NetFlow reports collected from a router into two subsets: `Large` and `Small`. Since the duration of a NetFlow report r is upper-bounded by the active timeout interval a, the flow needs to have sent more than H bytes within the a-sec period following the first-packet arrival time (f_r) for the NetFlow report r to qualify as `Large`.

Types of flows: We define a β flow as a flow that has only `Small` NetFlow reports, and a γ flow as a flow that has at least one `Large` NetFlow report. Thus, a γ flow may have multiple `Large` and `Small` NetFlow reports. Since TCP varies its sending rate, not all NetFlow reports of a γ flow will exceed the size threshold H. An α flow is defined to be a γ flow whose size and rate exceed specified (configurable) thresholds.

C. Algorithm

Using the notation in Table I, the main steps of the algorithm are listed below:

1) From each day's set of NetFlow reports, $\mathbf{F}_i$, determine sets $\mathbf{A}_i$, $\mathbf{W}_i$, and $\mathbf{B}_i$ using the size threshold H.
2) For each day i, the set $\mathbf{A}_i \bigcup \mathbf{B}_i$ is divided into disjoint subsets, $\mathbf{C}_{ij}$, $1 \leq j \leq |\mathbf{W}_i|$.
3) Order the reports in each set $\mathbf{C}_{ij}$ by sorting on the first-packet timestamp (earliest-to-latest). The ordered set of reports are $r_1, r_2, \cdots, r_{|\mathbf{C}_{ij}|}$.

TABLE I: Notation

i	per-day index
j	flow-identifier (ID) index
k	γ-flow index
r	NetFlow-report index
$\mathbf{F}_i$	set of NetFlow reports
$\mathbf{A}_i$	set of `Large` NetFlow reports (size $> H$)
$\mathbf{W}_i$	set of unique flow IDs ω_r for reports $r \in \mathbf{A}_i$
$\mathbf{B}_i$	set of `Small` NetFlow reports r whose flow IDs $\omega_r \in \mathbf{W}_i$
$\mathbf{C}_{ij}$	set of NetFlow reports r, s.t. $\omega_r = j$, for $j \in \mathbf{W}_i$
$\mathbf{E}_{ijk}$	Subset of $\mathbf{C}_{ij}$: reports of a single γ flow
N_{ij}	Number of γ flows
S_{ijk}	Size of γ flow
D_{ijk}	Duration of γ flow
ρ	packet sampling rate (e.g., 1/1000)

4) Divide each set $\mathbf{C}_{ij}$ into disjoint subsets $\mathbf{E}_{ijk}$, $1 \leq k \leq N_{ij}$ such that a consecutive set of NetFlow reports $\{r_n, r_{n+1}, \cdots, r_{n+u}\} \in \mathbf{E}_{ijk}$ iff

$$\begin{aligned} f_{r_{m+1}} - l_{r_m} &\leq \tau \quad n \leq m < n+u \\ f_{r_n} - l_{r_{n-1}} &> \tau \quad \text{for } n \neq 1 \\ f_{r_{n+u+1}} - l_{r_{n+u}} &> \tau \quad \text{for } n+u \neq |\mathbf{C}_{ij}| \end{aligned} \tag{2}$$

5) A γ flow $k, 1 \leq k \leq N_{ij}$, appearing on day i with flow-ID $\omega_j \in \mathbf{W}_i$, and consisting of NetFlow reports $\{r_n \cdots, r_{n+u}\} \in \mathbf{E}_{ijk}$, is characterized by

$$\begin{aligned} \text{Size } S_{ijk} &= (\frac{1}{\rho}) \sum_{m \in \mathbf{E}_{ijk}} o_m \\ \text{Duration } D_{ijk} &= l_{r_{n+u}} - f_{r_n} \\ \text{Av. rate } R_{ijk} &= \frac{S_{ijk}}{D_{ijk}} \end{aligned} \tag{3}$$

Starting with each day's set of NetFlow reports ($\mathbf{F}_i$), the *first step* is to find the subset of `Large` NetFlow reports ($\mathbf{A}_i$), from which the set of unique γ flow IDs ($\mathbf{W}_i$) is extracted. Using these flow IDs, a second pass through set $\mathbf{F}_i$ is executed to find all `Small` NetFlow reports (set $\mathbf{B}_i$) for the γ flows observed on day i. The goal of this first step is to reduce the number of NetFlow reports from which to extract α flows.

The *second step* create sets $\mathbf{C}_{ij}$ consisting of all the `Large` and `Small` NetFlow reports corresponding to each γ flow ID j. Since these $\mathbf{C}_{ij}$ sets are extracted from the disjoint sets of `Large` ($\mathbf{A}_i$) and `Small` ($\mathbf{B}_i$) NetFlow reports, the reports in each $\mathbf{C}_{ij}$ need to be sorted by the first-packet timestamp before flows can be reconstructed. This is the *third step*.

The *fourth step* is to divide the NetFlow reports in each set $\mathbf{C}_{ij}$ into multiple subsets, each of which consists of a set of consecutive reports belonging to a single γ flow. Recall from Section III-B, that if a time gap threshold is exceeded between the last-packet timestamp l_r of one NetFlow report r and the first-packet timestamp f_{r+1} of the next NetFlow report $(r+1)$,

the flow is considered to have terminated with report r, and a new flow begun with the next report. There is potential for a small gap between l_r and f_{r+1} for two consecutive reports r and $(r+1)$ because of packet sampling. Therefore, as long as this gap is less than a time-threshold τ, the consecutive NetFlow reports are considered to belong to the same flow. Using k as the index for γ flows, the subsets of $\mathbf{C}_{ij}$ are denoted $\mathbf{E}_{ijk}$, all of which share the same flow ID j in their appearance on day i (see Table I).

The *final step* is to add up the bytes in the NetFlow reports of each γ flow to determine the size of the flow and multiply by the reciprocal of the packet sampling rate ρ. Duration is computed by finding the time difference between the last-packet timestamp of the last NetFlow report and the first-packet timestamp of the first NetFlow report in each set $\mathbf{E}_{ijk}$. Average rate is computed by dividing flow size by flow duration.

As an example, consider the NetFlow reports shown in Table II. The first two columns show the number of packets, and cumulative number of bytes, in the sampled packets of the NetFlow report. The next five columns, source and destination IP addresses, source and destination transport-layer port numbers, and protocol type field, constitute the flow ID ω (see (1)). The source and destination IP addresses were anonymized and hence the numbers shown in Table II are not in the expected 4-byte format. The timestamps (TS) are in UTC format. For example, the first-packet TS of the first NetFlow report is 1.304269790137E9; UTC time 1304269790 corresponds to Sun, 01 May 2011 17:09:50 GMT [19]. The last three digits 137 corresponds to milliseconds. In this example, τ was set to 60 sec. The gap between the last-packet TS of the first NetFlow report and the first-packet TS of the next NetFlow report is 889.798 sec; as this gap is greater than τ (1 min), the second NetFlow report of Table II represents the start of a new flow. This flow had $(95 + 6 = 101)$ sequential NetFlow reports with inter-report gaps less than τ. For example, the gap between the first two reports of the 101-report flow is only 180 ms. Similarly, the gap between the last-packet TS of the last report of the 101-report flow and the first-packet TS of the last report in Table II is 40665.873 sec, which is well above τ.

IV. Validation of algorithm

A. Method

To validate the algorithm presented in Section III-C, we devised the following method using operational, not experimental, datasets.

Step 1: Obtain GridFTP usage logs from an operational data transfer node: GridFTP usage logs were obtained from dedicated data transfer nodes at the National Energy Research Scientific Computing (NERSC) center for the period, Apr. 22 to June 30, 2012. The usage logs include the following information for each transfer: remote end's IP address, size in bytes, start time of the transfer, and transfer duration.

Step 2: Find corresponding NetFlow reports from an ESnet router: Next, since NERSC is a customer of ESnet, and ESnet has located one of its routers at NERSC, i.e., a provider-edge (PE) router, we obtained NetFlow reports from this PE router for the same time period. For each GridFTP usage log entry, using the source and destination IP addresses and the start and end time of the corresponding transfer, our software finds matching NetFlow reports.

Step 3: Find additional NetFlow reports with the same flow IDs: Using the unique 5-tuple flow IDs from the per-day set of matched NetFlow reports obtained in Step 2, a second pass was executed to find all NetFlow reports corresponding to these 5-tuple flow IDs even if the time intervals of these reports (first-packet TS, last-packet TS) were outside any GridFTP-transfer time intervals. These NetFlow reports were required to determine whether our size/rate estimation algorithm could correctly identify the GridFTP transfers as single flows.

Step 4: Characterize flows: From the sets of NetFlow reports found in steps 2 and 3, we executed the algorithm described in Section III-C to characterize γ flows.

Step 5: Recreate "sessions" from GridFTP transfer logs: Our prior analysis [6] showed that most GridFTP transfers occur in sessions, i.e., multiple file transfers on the same TCP connection. The `-fast` option of GridFTP when invoked to move files in a directory will result in all files being transferred on the same TCP connection. The GridFTP sending process sends multiple files concurrently. All transfers to the same destination with overlapping durations are included in a single session. A gap value of up to 10 ms was allowed when grouping transfers into sessions. Also, the log entry shows the number of parallel TCP streams used for a transfer (which is set by users with the `-p` option). Since large datasets are typically moved using the `-p` option, we included only those transfers that used more than 1 parallel TCP stream. All transfers within each session had the same number of parallel TCP streams.

Step 6: Accuracy computation: For each GridFTP session that exceeded size and rate thresholds (5 GB and 667 Mbps), we found multiple γ flows whose start and end times fell within the GridFTP session duration. There were multiple γ flows because of the use of parallel TCP streams. The γ-flow sizes were added to find the total size before comparing with the GridFTP session size. The average duration across all the γ flows corresponding to a GridFTP session was determined and compared with the GridFTP session duration. Size (duration) accuracy is defined as the ratio of the size (duration) estimated by our algorithm from the NetFlow reports to the size (duration) reported in the GridFTP usage logs.

B. Results

Table III shows the results of our validation procedure. Both duration accuracy and size accuracy for these high-rate large-sized flows were close to 100%. Size accuracy can be greater than 100% because the NetFlow packet sampling process

TABLE II: Example NetFlow reports observed for one γ flow ID in one day; TS: Timestamp; dur: duration (sec)

pkts	bytes	src IP	dst IP	src port	dst port	prot.	first-pkt TS	last-pkt TS	dur (sec)
Previous flow's last NetFlow report									
481	683020	6853	6840	20886	62362	6	1.304269790137E9	1.304269820122E9	29.98
Next flow (has 101 NetFlow reports)									
173	245660	6853	6840	20886	62362	6	1.304270709920E9	1.304270749856E9	39.93
251	356420	6853	6840	20886	62362	6	1.304270750036E9	1.304270809975E9	59.93
247	350740	6853	6840	20886	62362	6	1.304270810282E9	1.304270869675E9	59.39
There were 95 other NetFlow reports with inter-report gaps less than τ									
230	326600	6853	6840	20886	62362	6	1.304276573971E9	1.304276633668E9	59.69
234	332280	6853	6840	20886	62362	6	1.304276634016E9	1.304276693903E9	59.88
61	86620	6853	6840	20886	62362	6	1.304276694116E9	1.304276704044E9	9.92
Next flow's first NetFlow report									
57	80940	6853	6840	20886	62362	6	1.304317369174E9	1.304317391838E9	22.66

TABLE III: Results of algorithm validation using GriFTP logs

No.	Log dur. (s)	Est. dur. (s)	D-acc (%)	Log size (GB)	Est. size (GB)	S-acc (%)
1	195.3	194.2	99.4	52.4	51.9	99.0
2	158.9	156.2	98.3	34.4	33.2	96.7
3	190.2	187.7	98.7	34.4	34.3	99.9
4	157.8	155.4	98.5	34.4	35	101.7
5	6516	6466.3	99.2	6.2	6.6	105.5
6	7696.8	7695.8	99.9	6.2	6.3	101.3
7	73.94	72	97.4	5.8	6.1	105.5

could have caught more packets of a particular transfer than 1-in-1000.

V. Application to ESnet traffic

We procured NetFlow reports from four ESnet routers for a 7-month time period, May-Nov. 2011, and applied the algorithm described in Section III-C to characterize γ and α flows. As the same data was used for a different analysis, the description about the routers is taken from a published paper [7]. Routers `router-1` and `router-2` are provider-edge (PE) routers located in ESnet customers' sites, and hence connected to a single customer, a DOE national laboratory (lab) network each. `Router-3` is a core router connected to multiple ESnet PE routers, and multiple national and international REN peers, such as Internet2 and AARnet. `Router-4` is one of the ESnet routers used for commercial peering. NetFlow data is collected only for the input direction of inter-domain interfaces. Since ESnet does not offer transit service, all packets are either sourced or destined to ESnet's customers (DOE labs). The NetFlow reports collected at `router-1` and `router-2` are for *downloads* executed from DOE lab machines, while NetFlow reports collected at `router-3` and `router-4` are for *uploads* to DOE lab machines.

After presenting the results generated by applying our algorithm to the NetFlow data in Section V-A, the implications of these findings are discussed in Section V-B.

A. Results

Four sets of results are presented:

1) aggregate characteristics of γ flows and α flows
2) statistics about three characteristics: size, rate, and duration, of γ flows and α flows
3) number of α flows as a function of the size and rate thresholds, and
4) persistency measure: number of γ flows and α flows created between the same source and destination.

Aggregate characteristics of γ flows (H was set to 1 GB) and α flows (using a size threshold of 5 GB and rate threshold of 100 Mbps) at each of the routers across the observation period of 214 days are listed in Table IV. The second row shows the number of unique γ flow IDs observed, while the third row lists the number of unique source-destination pairs that generated γ flows, in the 214-day period. The fourth row represents the maximum number of per-day γ flows corresponding to a single γ flow ID. Multiple γ flows could have resulted from a TCP connection being held open for a long duration with gaps between flows as explained in Section III-B. The last three rows present aggregate information about α flows.

Statistics for three characteristics of γ flows: size, rate, and duration, are presented in Tables V, VI, and VII. These tables are independent, e.g., the largest-sized flow is not the same as the highest-rate flow.

Table VIII presents results from a sensitivity analysis of the number of α flows to the size and rate thresholds.

Finally, we characterized the persistency with which source-destination pairs generated γ flows and α flows. Figs. 1 and 2 plot the cumulative distribution function (CDF) of the numbers of γ flows and α flows per source/destination pair for `router-2`, `router-3` and `router-4`. The plots for `router-1` have been omitted because they overlapped significantly with those of `router-2`. Recall that `router-1` and `router-2` are PE routers that capture flows corresponding to downloads from DOE labs, and hence have similar numbers of flows.

TABLE IV: Aggregate data on γ and α flows; across 214 days

	Routers, router–			
	1	2	3	4
No. of γ flows	28685	27963	2516	212
No. of unique γ flow IDs	19365	26939	2455	212
No. of unique /32 src-dst pairs gen. γ flows	1479	1611	193	158
Max. no. of per-day γ flows corr. to a single γ flow ID	33	56	6	1
No. of α flows	916	9538	986	16
No. of unique α flow IDs	834	9043	943	16
No. of unique /32 src-dst pairs gen. α flows	95	419	89	14

TABLE V: Size in MB of γ flows; across 214 days

	Routers, router–			
	1	2	3	4
Min	1001	1001	1005	1010
1st Qu.	1149	1540	4050	1203
Median	1275	2869	4360	1532
Mean	2513	9046	17540	3612
3rd Qu	1701	8768	21380	3772
90%	2761	16600	54115	5774
99%	12909	92012	104356	26389
99.9%	229727	288797	180138	100460
Max	633300	811600	233600	112800
CV	5.20	2.56	1.40	2.43
skewness	25.35	12.56	2.37	10.09

TABLE VI: Rate in Mbps of γ flows; across 214 days

	Routers, router–			
	1	2	3	4
Min	11.7	3.6	34.6	49.2
1st Qu.	160.9	147	117.6	130.9
Median	199.3	181.9	132.6	156.4
Mean	245.2	230.9	159	182.7
3rd Qu.	258.9	252.1	159.2	195.8
90%	403	363	264	275
99%	881	944	503	649
99.9%	1711	993	953	755
Max	5154	5757	979	776
CV	0.71	0.72	0.56	0.61
skewness	7.36	3.95	3.82	2.86

B. Discussion

The results presented in the previous section are discussed below in three groupings. *First*, we discuss the numerical values themselves to understand the range of sizes, rates, durations, and frequencies, of γ flows and α flows. *Next*, we compare the characteristics of flows observed at the different routers. *Finally*, an example application is described to demonstrate usage of this characterization of α flows.

TABLE VII: Duration in sec of γ flows; across 214 days

	Routers, router–			
	1	2	3	4
Min	4.2	8.04	9.5	12
1st Qu.	41.8	60.9	190.9	54.9
Median	54.2	121.1	272	94.3
Mean	122.8	414.2	1098	235.6
3rd Qu.	73.58	398.9	1169	227.6
Max	32460	31910	13940	9978
CV	7.39	2.34	1.50	3.18
skewness	23.76	10.33	2.32	10.99

TABLE VIII: Sensitivity to size-rate threshold: No. of α flows

size	rate	Routers, router–			
		1	2	3	4
5GB	200Mbps	496	4475	201	3
10GB	100Mbps	526	5460	726	3
10GB	150Mbps	399	4121	297	1
10GB	180Mbps	375	3037	124	0
10GB	200Mbps	357	2443	92	0
50GB	200Mbps	19	505	28	0
80GB	500Mbps	0	20	0	0

Numerical values:

The difference between the number of γ flows, and number of unique γ flow IDs (rows 1 and 2 in Table IV) occurs because of two possibilities: the same 5-tuple values were used on two

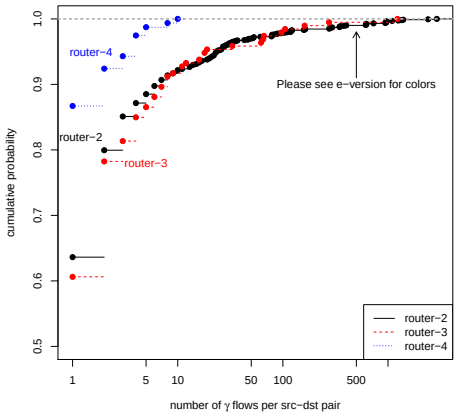

Fig. 1: CDF of number of γ flows per src/dst pair across 214 days for router−2, router−3, router−4 (router−1 plot overlaps closely with the router−2 plot and is hence omitted)

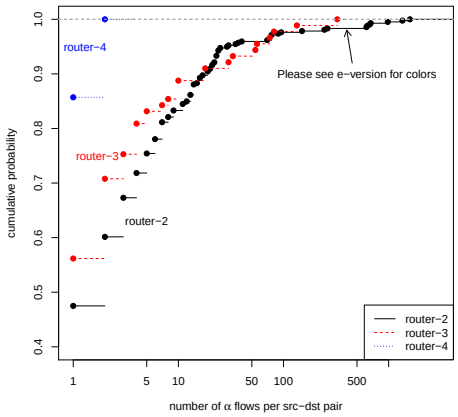

Fig. 2: CDF of number of α flows ($>$ 5 GB, $>$ 100 Mbps) per src/dst pair across 214 days for `router-2`, `router-3`, `router-4`

different days, or a given flow ID was reused in multiple flows within the same day. The latter is characterized in the fourth row. Most γ flow IDs have only single γ flows in a given day, but there are a few occasions when multiple γ flows have been observed on the same day for a given γ flow ID. As many as 56 γ flows were observed for a single five-tuple ID in one day (at `router-2`) as shown in Table IV.

Across the 214-day period, of all the flows observed at the four routers, the largest-sized flow was 811.6 GB (max row of Table V) and the highest-rate flow enjoyed an aggregate rate of 5.76 Gbps (max row of Table VI), both of which were downloads passing through PE router `router-2`. The largest-sized flow had a rate of 301 Mbps, and the fastest flow size was 7.14 GB. The longest flow lasted 32460 sec (more than 9 hours) passing through `router-1`, during which time 370 GB was moved (max row of Table VII).

At the lower end, rates as low as 3.6 Mbps were observed, also at `router-2`. This particular γ flow moved 1.9 GB, which means it lasted about 4181 sec (more than an hour).

Since there is a significant gap between the 3^{rd} quartile values, and the maximum values, Tables V and VI show a few more quantiles in the fourth quarter. Using the number of γ flows provided in Table IV, we see that the 99.9% value of 229.73 GB implies that only 28 flows in the size range (229.73 GB, 633.3 GB) entered `router-1` from its connected DOE lab. Similarly, the 99.9% rate value for γ flows passing through `router-2` was still less than 1 Gbps (even though the maximum rate for this router was 5.76 Gbps). This implies that only 27 flows out of the 27963 observed γ flows (flows larger than 1 GB with a rate $>$ 133 Mbps) enjoyed (average) rates higher than 1 Gbps during the 7-month period.

Skewness is defined as μ_3/σ^3, where μ_3 is the third moment and σ is the standard deviation. The coefficient of variation (CV) and skewness values were lower for rates than for sizes, as seen in Tables V and VI. This was expected since file sizes have heavy-tailed distributions [20].

Table VIII shows that the number of α flows falls quickly as the size-rate threshold is increased, which is to be expected. Nevertheless, the absolute numbers are interesting to note. Router `router-2` connects ESnet to a supercomputing center, which explains that even at the high per-flow thresholds of 80 GB and 500 Mbps, 20 α flows were observed.

Comparison between flows observed at different routers:

As seen in Table IV, there were many more γ flows in downloads from DOE labs than uploads to DOE labs (since downloads were observed at `router-1` and `router-2`, while uploads were observed at `router-3` and `router-4`). Also, more source-destination pairs engaged in transfers larger than 1 GB for downloads than uploads.

As seen in Tables V and VI, γ flows for downloads from DOE labs were larger in size and higher in rate. Uploads to DOE labs, observed at `router-3` and `router-4` were considerably slower, with the maximum rate reaching only 776 Mbps at the commercial peering router `router-4` and only 979 Mbps at the REN-peering router `router-3`. Maximum flow sizes were also smaller. Table VII shows that the longest downloads were longer than the longest uploads, but most γ flows are short in duration.

A comparison of the number of α flows across the 4 routers from Table VIII shows a difference between the two PE routers. While `router-1` is a PE router connected to large national DOE lab, the significant research projects at this lab are in a single science discipline. In contrast, PE router `router-2` connects to a national scientific supercomputing center that is used by scientists from many disciplines. This explains the larger numbers of α flows for `router-2` when compared to `router-1` as seen in Table VIII.

Finally, Figs. 1 and 2 show that uploads through the commercial peering router `router-4` were considerably fewer (maximum values of 10 γ flows and 2 α flows) than through the other routers. A comparison of the red (`router-3`) and black (`router-2`) plots shows the former plots ending before the latter plots. The maximum number of γ-flow and α-flow uploads per source-destination pair for `router-3` were 1229 and 325, respectively, while at `router-2`, the numbers for γ-flow and α-flow downloads per source-destination pair were 2913 and 1596, respectively. The maximum γ-flow and α-flow downloads per source-destination pair at `router-1` were 2860 and 445, respectively. The ninety percentile numbers for γ flows per source-destination pair were 39, 7, 7.8 and 2 for the four routers in sequence, and the numbers for α flows per source-destination pair were 11.6, 19, 18 and 1.7. Therefore, less than 10% of the source-destination pairs generated large numbers of repeated γ flows and α flows, which makes it somewhat easier for operators to provide better services (higher rates, lower variance) for these particular source-destination pairs.

Example application:

Consider the source-destination pair that generated the largest numbers of γ flows and also the largest number of α flows across the 214-day period. The particular source-

destination IP address pair that generated these maximum number of flows was (2888,7128) using the anonymized addresses[3]. Since all these flows were between the same source and destination, and there were no network upgrades during the data-collection period, the bottleneck link rate and round-trip time were approximately the same, and all flow sizes are greater than 1 GB, which means TCP's Slow Start period could not have had a major influence on the average rate. Nevertheless, in the 2913 γ-flow set, 75% of the flows experienced less than 161.2 Mbps while the highest rate experienced was 1.1 Gbps (size: 3.5 GB). Similarly, in the 1596 α-flow set, 75% of the flows experienced less than 167 Mbps, while the highest rate experienced was 536 Mbps (size: 11 GB). Such information would allow the provider to initiate diagnostics to determine the causes of lower rates.

VI. Conclusions

This work demonstrated that it is feasible to determine the size, duration, and rate, of high-rate, large-sized (α) flows from NetFlow reports in spite of low packet sampling rates, e.g., 1-in-1000. The algorithm proposed here can form the basis of a network management system for characterizing α flows. Example applications include special traffic-engineering of α flows (since they have the potential to degrade service quality of real-time flows), offering users who generate α flows diagnostic support to determine causes of low throughput or high throughput variance, and identifying BGP misconfigurations that cause α flows to enter a provider's network on a less-preferred route. The algorithm was validated using independently collected usage logs from application servers. We executed our algorithm on actual NetFlow reports from 4 ESnet routers collected over a 7-month period. Individual flows moving datasets as large as 811 GB and at rates as high as 5.7 Gbps were observed. Some source-destination pairs were found to repeatedly create α flows. An analysis of the rates of the 1596 repeated α flows created by one pair showed considerable variance, with minimum rate of 100 Mbps, maximum rate of 536 Mbps, and a coefficient of variation of 30%.

VII. Acknowledgment

The University of Virginia portion of this work was supported by the U.S. Department of Energy (DOE) grant DE-SC0007341 and NSF grants OCI-1038058, OCI-1127340, and CNS-1116081. The ESnet portion of this work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. DOE under Contract No. DE-AC02-05CH11231.

References

[1] S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level analysis and modeling of nework traffic," in *ACM SIGCOMM Internet Measurement Workshop 2001*, November 2001, pp. 99–104.

[2] ESnet. [Online]. Available: http://www.es.net/

[3] Internet2. [Online]. Available: http://www.internet2.edu/

[4] "Terabit networks for extreme-scale science workshop report." [Online]. Available: http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Terabit_networks_workshop_report.pdf

[5] GridFTP. [Online]. Available: http://globus.org/toolkit/docs/3.2/gridftp/

[6] Z. Liu, M. Veeraraghavan, Z. Yan, C. Tracy, J. Tie, I. Foster, J. Dennis, J. Hick, Y. Li, and W. Yang, "On using virtual circuits for GridFTP transfers," in *The International Conference for High Performance Computing, Networking, Storage and Analysis 2012 (SC 2012)*, Nov. 10-16, 2012, pp. 81:1–81:11.

[7] T. Jin, C. Tracy, M. Veeraraghavan, and Z. Yan, "Traffic engineering of high-rate large-sized flows," in *Proc. of IEEE 14th High Performance Switching and Routing (HPSR) 2013*, July 8-11 2013.

[8] perfSONAR. [Online]. Available: http://www.perfsonar.net/

[9] N. Kamiyama and T. Mori, "Simple and accurate identification of high-rate flows by packet sampling," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, 2006, pp. 1–13.

[10] T. Mori, M. Uchida, R. Kawahara, J. Pan, and S. Goto, "Identifying elephant flows through periodically sampled packets," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, ser. IMC '04. New York, NY, USA: ACM, 2004, pp. 115–120. [Online]. Available: http://doi.acm.org/10.1145/1028788.1028803

[11] Y. Zhang, B. Fang, and Y. Zhang, "Identifying high-rate flows based on bayesian single sampling," in *2010 2nd International Conference on Computer Engineering and Technology (ICCET)*, vol. 1, 2010, pp. V1–370–V1–374.

[12] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," *IEEE/ACM Transactions on Networking*, vol. 13, no. 5, pp. 933–946, 2005.

[13] Y. Lu, M. Wang, B. Prabhakar, and F. Bonomi, "ElephantTrap: A low cost device for identifying large flows," in *15th Annual IEEE Symposium on High-Performance Interconnects, 2007. HOTI 2007.*, 2007, pp. 99–108.

[14] M. Kodialam, T. V. Lakshman, and S. Mohanty, "Runs based traffic estimator (rate): a simple, memory efficient scheme for per-flow rate estimation," in *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2004, pp. 1808–1818 vol.3.

[15] F. Hao, M. Kodialam, T. V. Lakshman, and H. Zhang, "Fast, memory-efficient traffic estimation by coincidence counting," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 3, 2005, pp. 2080–2090 vol. 3.

[16] M. Zadnik, M. Canini, A. Moore, D. Miller, and W. Li, "Tracking elephant flows in Internet backbone traffic with an FPGA-based cache," in *International Conference on Field Programmable Logic and Applications, 2009. FPL 2009.*, 2009, pp. 640–644.

[17] J. Paisley and J. Sventek, "Real-time detection of grid bulk transfer traffic," in *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, april 2006, pp. 66 –72.

[18] N. Hohn and D. Veitch, "Inverting sampled traffic," *IEEE/ACM Transactions on Networking*, vol. 14, no. 1, pp. 68–80, 2006.

[19] Epoch & Unix Timestamp Conversion Tools. [Online]. Available: http://www.epochconverter.com/

[20] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *IEEE/ACM Transaction on Networking*, vol. 3, pp. 226–244, 1995.

[3]For privacy reasons, the actual addresses are not published, but are stored in our data archives for retrieval if needed.